

## Chapitre 2. Le modèle de régression multiple

Le modèle de régression multiple est une extension du modèle de régression simple. Dans ce présent chapitre, il sera présenté le modèle linéaire général, la procédure d'estimation des paramètres en étudiant les propriétés statistiques des estimateurs, les différents tests d'hypothèses concernant les coefficients du modèle et l'analyse de la variance ainsi qu'aux tests s'y rattachant.

### I. Le modèle linéaire général

Dans la régression simple, nous avons considéré qu'une variable endogène est expliquée à l'aide d'une seule variable exogène. Cependant, il est extrêmement rare qu'un phénomène économique ou social puisse être appréhendé par une seule variable. Le modèle linéaire général est une généralisation du modèle de régression simple dans lequel figurent plusieurs variables explicatives :

$$y_t = a_0 + a_1x_{1t} + a_2x_{2t} + \dots + a_kx_{kt} + \varepsilon_t \quad \text{pour } t=1, \dots, n \quad \text{avec :}$$

$y_t$  = variable à expliquer à la date  $t$ ;

$x_{1t}$  = variable explicative 1 à la date  $t$ ;

$x_{2t}$  = variable explicative 2 à la date  $t$ ;

$\dots x_{kt}$  = variable explicative  $k$  à la date  $t$ ;

$a_0, a_1, \dots, a_k$  = paramètres du modèle ;

$\varepsilon_t$  = erreur de spécification (différence entre le modèle vrai et le modèle spécifié), cette erreur est inconnue et restera inconnue;

$n$  = nombre d'observations.

L'écriture précédente du modèle est d'un maniement peu pratique. Afin d'en alléger l'écriture et de faciliter l'expression de certains résultats, on a habituellement recours aux notations matricielles.

En écrivant le modèle, observation par observation, nous obtenons :

$$y_1 = a_0 + a_1x_{11} + a_2x_{21} + \dots + a_kx_{k1} + \varepsilon_1$$

$$y_2 = a_0 + a_1x_{12} + a_2x_{22} + \dots + a_kx_{k2} + \varepsilon_2$$

.....

$$y_t = a_0 + a_1x_{1t} + a_2x_{2t} + \dots + a_kx_{kt} + \varepsilon_t$$

.....

$$y_n = a_0 + a_1x_{1n} + a_2x_{2n} + \dots + a_kx_{kn} + \varepsilon_n$$

$$\text{Soit, sous forme matricielle : } \underset{(n,1)}{Y} = \underset{(n, k+1)}{X} \underset{(k+1, 1)}{a} + \underset{(n,1)}{\varepsilon} \quad (1)$$

Avec :

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_t \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}; a = \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_t \\ \vdots \\ a_k \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_t \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

La première colonne de la matrice X, composée de 1, qui correspond au coefficient  $a_0$  (coefficient du terme constant). La dimension de la matrice X est donc de  $n$  lignes et  $k+1$  colonnes ( $k$  étant le nombre de variables explicatives réelles, c'est-à-dire constante exclue).

## II. Estimation et propriétés des estimateurs

### 2.1 . Estimation des coefficients de régression

Soit le modèle sous forme matricielle à  $k$  variables explicatives et  $n$  observations :

$$Y = Xa + \varepsilon \quad (1)$$

Afin d'estimer le vecteur  $a$  composé des coefficients  $a_0, a_1, \dots, a_k$ , nous appliquons la **méthode des Moindres Carrés Ordinaires (MCO)** qui consiste à **minimiser la somme des carrés des erreurs**, soit :

$$\text{Min} \sum_{t=1}^n \varepsilon_t^2 = \text{Min} \varepsilon' \varepsilon = \text{Min} (Y - Xa)'(Y - Xa) = \text{Min} S \quad (2)$$

avec  $\varepsilon'$  transposé du vecteur  $\varepsilon$ .

Pour minimiser cette fonction par rapport à  $a$ ,  $S$  sera différenciée par rapport à  $a$

$$\frac{\partial S}{\partial a} = -2 X'Y + 2 X'X\hat{a} = 0 \rightarrow \hat{a} = (X'X)^{-1} X'Y \quad (3)$$

Cette solution est réalisable si la matrice carrée  $X'X$  de dimension  $(k+1, k+1)$  est inversible. La matrice  $X'X$  est la matrice des produits croisés des variables explicatives ; en cas de colinéarité parfaite entre deux variables explicatives, la matrice  $X'X$  est singulière et la méthode des MCO défaille.

On appelle équations normales les équations issues de la relation :

$$X' X \hat{a} = X' Y$$

Le modèle estimé s'écrit :

$$\hat{y}_t = \hat{a}_0 + \hat{a}_1 x_{1t} + \hat{a}_2 x_{2t} + \cdots + \hat{a}_k x_{kt} + e_t$$

### Effet de la variation d'une seule des variables explicatives

Si la variable  $x_2$  passe de la valeur  $x_{2t}$  à  $(x_{2t} + \Delta x_{2t})$ , toutes choses étant égales par ailleurs (les  $k-1$  autres variables restant constantes), alors la variable à expliquer varie de  $\hat{a}_2 \Delta x_2$ .

Les coefficients s'interprètent donc directement en termes de propension marginale.

### 2.2 Hypothèses et propriétés des estimateurs

Par construction, le modèle est linéaire en X (ou sur ces coefficients) et nous distinguons les hypothèses stochastiques (liées à l'erreur  $\varepsilon$ ) des hypothèses structurelles.

#### 1) Hypothèses stochastiques

– H1 : les valeurs  $x_{i,t}$  sont observées sans erreur.

- H2 :  $E(\varepsilon_t)=0$ , l'espérance mathématique de l'erreur est nulle.
- H3 :  $E(\varepsilon_{2t})=\sigma_{2\varepsilon}$ , la variance de l'erreur est constante ( $\forall t$ )(homoscédasticité).
- H4 :  $E(\varepsilon_t \varepsilon_{t'})=0$  si  $t \neq t'$ , les erreurs sont non corrélées (ou encore indépendantes).
- H5 :  $Cov(x_{it}, \varepsilon_t)=0$ , l'erreur est indépendante des variables explicatives ( problème de l'endogénéité.)

## 2) Hypothèses structurelles

- H6 : absence de colinéarité entre les variables explicatives, cela implique que la matrice  $(X'X)$  est régulière et que la matrice inverse  $(X'X)^{-1}$  existe.
- H7:  $(X'X)/n$  tend vers une matrice finie non singulière.
- H8:  $n > k+1$ , le nombre d'observations est supérieur au nombre des séries explicatives.

## 3) Propriétés des estimateurs

L'estimateur est sans biais :  $E(\hat{a}) = a$

La matrice des variances et covariances de l'erreur  $\varepsilon$  est donnée comme suit :

$$\Omega_{\hat{a}} = \hat{\sigma}_{\varepsilon}^2 (X'X)^{-1}$$

L'estimateur est convergent

## 2.3. Équation d'analyse de la variance et qualité d'un ajustement

Comme pour le modèle de régression simple, l'équation fondamentale d'analyse de la variance est donnée comme suit :

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t e_t^2$$

$$\text{SCT} = \text{SCE} + \text{SCR}$$

*La variabilité totale (SCT) est égale à la variabilité expliquée (SCE) + la variabilité des résidus (SCR).*

Cette équation permet de **juger de la qualité de l'ajustement d'un modèle** ; en effet, plus la variance expliquée est « proche » de la variance totale, meilleur est l'ajustement global du modèle. C'est pourquoi nous calculons le rapport SCE sur SCT:

$$R^2 = \frac{\sum_t (\hat{y}_t - \bar{y})^2}{\sum_t (y_t - \bar{y})^2} = 1 - \frac{\sum_t e_t^2}{\sum_t (y_t - \bar{y})^2}$$

$R^2$  est appelé le coefficient de détermination, et R le coefficient de corrélation multiple.  $R^2$  mesure la proportion de la variance de Y expliquée par la régression de Y sur X. Cette qualité de l'ajustement et l'appréciation que l'on a du  $R^2$  doivent être tempérées par le degré de liberté de l'estimation. En effet, lorsque le degré de liberté est faible, il convient de corriger le  $R^2$  afin de tenir compte du relativement faible nombre d'observations comparé au nombre de facteurs explicatifs par le calcul d'un  $R^2$  « corrigé »

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} (1 - R^2)$$

## III. Les tests statistiques

### 3.1 Comparaison d'un paramètre $a_i$ à une valeur fixée a

Le test d'hypothèses est le suivant

$$H_0 : a_i = \bar{a}$$

$$H_0 : a_i \neq \bar{a}$$

Nous savons que :  $\frac{\hat{a}_i - a_i}{\hat{\sigma}_{\hat{a}_i}}$  suit donc une loi de Student à  $(n-k-1)$  degrés de liberté

Sous l'hypothèse  $H_0$ , cette relation devient :  $\frac{|\hat{a}_i - \bar{a}|}{\hat{\sigma}_{\hat{a}_i}} = t_{\hat{a}_i}^* \rightarrow$  loi de Student à  $(n-k-1)$  degrés de liberté.

Si  $t_{\hat{a}_i}^* > t^{\alpha/2}_{(n-k-1)}$  alors nous rejetons l'hypothèse  $H_0$ ,  $a_i$  est significativement différent de  $a$  (au seuil de  $\alpha$ ).

Si  $t_{\hat{a}_i}^* < t^{\alpha/2}_{(n-k-1)}$  alors nous acceptons l'hypothèse  $H_0$ ,  $a_i$  n'est pas significativement différent de  $a$  (au seuil de  $\alpha$ ).

**Cas particulier** : test par rapport à une valeur particulière  $a=0$ .

Si nous désirons savoir si une variable explicative figurant dans un modèle est réellement – significativement – contributive pour expliquer la variable endogène, il convient de tester si son coefficient de régression est significativement différent de 0 pour un seuil choisi, en général  $\alpha=5\%$ .

Sous  $H_0$  ( $a_i=0$ ), devient :

$\left| \frac{\hat{a}_i}{\hat{\sigma}_{\hat{a}_i}} \right| = t_{\hat{a}_i}^* \rightarrow$  loi de Student à  $(n-k-2)$  degrés de liberté.

$t_{\hat{a}_i}^*$  est appelé le ratio de Student, les règles de décision citées plus haut s'appliquent alors.

Ce test est très important ; en effet, si dans un modèle estimé, un des coefficients (hormis le terme constant) n'est pas significativement différent de 0, il convient d'éliminer cette variable et de ré-estimer les coefficients du modèle. La cause de cette non-significativité, est due :

- soit à une absence de corrélation avec la variable à expliquer,
- soit à une colinéarité trop élevée avec une des variables explicatives.

#### IV. L'analyse de la variance

##### A. Construction du tableau d'analyse de la variance et test de signification globale d'une régression

Dans cette section, nous allons nous interroger sur la signification globale du modèle de régression, c'est-à-dire si l'ensemble des variables explicatives a une influence sur la variable à expliquer. Ce test peut être formulé de la manière suivante : existe-t-il au moins une variable explicative significative ? Soit le test d'hypothèses :

$H_0 : a_1 = a_2 = \dots = a_k = 0$  (tous les coefficients sont nuls)

$H_1$  : il existe au moins un des coefficients non nul

Nous ne testons pas le cas où le terme constant  $a_0$  est nul, car seules nous intéressent les variables explicatives. Un modèle dans lequel seul le terme constant est significatif n'a aucun sens économique.

Le cas où l'hypothèse  $H_0$  est acceptée signifie qu'il n'existe aucune relation linéaire significative entre la variable à expliquer et les variables explicatives (ou encore que la Somme des Carrés Expliqués n'est pas significativement différente de 0).

D'après l'équation fondamentale d'analyse de la variance, la régression est jugée significative si la variabilité expliquée est significativement différente de 0. Le tableau 1 présente le tableau d'analyse de la variance permettant d'effectuer le test de Fisher.

$$\text{Ou encore : } F^* = \frac{\sum_t (\hat{y}_t - \bar{y})^2 / k}{\sum_t e_t^2 / (n-k-1)} = 1 - \frac{R^2 / k}{(1-R^2) / (n-k-1)}$$

Tableau 1 – Analyse de la variance pour une régression multiple

Source de variation	Somme des carrés	Degré de liberté	Carées moyen
$x_1, x_2, \dots, x_{1k}$	$SCE = \sum_t (\hat{y}_t - \bar{y})^2$	k	SCE/k
Résidu	$SCR = \sum_t e_t^2$	n-k-1	SCR/n-k-1
Total	$SCT = \sum_t (y_t - \bar{y})^2$	n-1	

L'hypothèse de normalité des erreurs implique que sous l'hypothèse H0, F\* suit une loi de Fisher (rapport de deux chi-deux). Nous comparons ce F\* calculé au F théorique à k et (n-k-1) degrés de liberté : si F\* > F nous rejetons l'hypothèse H0, le modèle est globalement explicatif.

## V. L'utilisation de variables indicatrices

Une variable indicatrice<sup>1</sup> est une variable explicative particulière qui n'est composée que de 0 ou de 1. Cette variable est utilisée lorsque, dans un modèle, nous désirons intégrer un facteur explicatif binaire : « le phénomène a lieu ou n'a pas lieu » pour corriger, par exemple, d'une valeur anormale ; ou bien lorsque le facteur explicatif est qualitatif : « le genre d'un individu, homme ou femme ». Il s'agit donc d'incorporer une ou des variables explicatives supplémentaires au modèle spécifié initialement et d'appliquer les méthodes classiques d'estimation. Le modèle de régression diffère selon l'apparition du phénomène par les valeurs d'un ou plusieurs coefficients alors que les autres paramètres sont identiques. En cas de modification structurelle d'un coefficient de régression, la variable muette affecte alors le coefficient de la ou des variables explicatives considérées.

Le domaine d'utilisation des variables indicatrices est très vaste, nous pouvons citer : la correction des valeurs anormales, la modification structurelle (0 pour la période avant le changement structurel, 1 après le changement structurel), l'intégration de la saisonnalité, la caractérisation d'un individu (genre, situation matrimoniale...), l'intégration de facteurs qualitatifs (appartenance d'un pays à la zone euro, promotion non quantifiable...), etc.

## V : Problèmes associés à l'analyse de régression

1. La relation qui unit la variable dépendante Y aux variables X est linéaire (les variables X et Y n'ont pas besoin d'être linéaires)
2. Le terme d'erreur est une variable aléatoire distribuée normalement avec une moyenne de 0 et une variance  $\sigma^2$  constante.

<sup>1</sup> Les termes de variables indicatrices, de variables auxiliaires ou de variables muettes sont indifféremment employés en français. Le terme anglo-saxon *dummy* est le plus couramment utilisé.

- Variance pas constante → Hétéroscédasticité
3.  $Cov(X, \varepsilon) = 0$ 
    - Termes d'erreurs corrélés avec les variables explicatives → Endogénéité
  4.  $Cov(\varepsilon_i, \varepsilon_j) = 0$ 
    - Termes d'erreurs corrélés entre eux → Autocorrélation
  5.  $Cov(X_j, X_k) = 0$ 
    - Variables indépendantes corrélées → Multicollinéarité

## 1- l' hétéroscédasticité

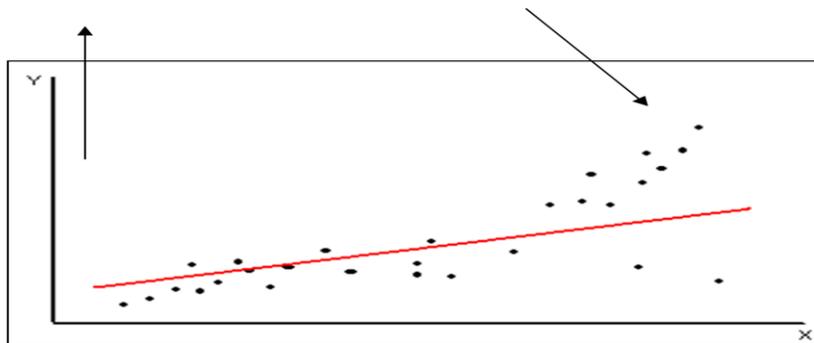
Le problème d'hétéroscédasticité est fréquent dans le cas d'une série de données en coupe transversale, notamment dans les données microéconomiques avec un effet de taille (l'existence d'une différence en termes de grandeur entre les observations).

La variance des termes d'erreur diffère entre les observations

Les paramètres estimés demeurent valides

Les tests de signification du modèle (*F*-test et *t*-test) ne sont pas fiables

La dispersion des résidus n'est pas constante



Dans l'estimation MCO les erreurs sont supposées être homoscedastiques, c'est-à-dire que la variance de l'erreur est constante pour l'ensemble des observations. Lorsque cette hypothèse n'est pas satisfaite l'estimation par MCO n'est pas optimale. Tandis que dans un modèle avec hétéroscédasticité la variance de l'erreur n'est pas constante d'une observation à l'autre, la matrice variance-covariance s'écrit comme suit :

$$\Omega_{\varepsilon} = \begin{pmatrix} E(\varepsilon_1 \varepsilon_1) & E(\varepsilon_1 \varepsilon_2) & E(\varepsilon_1 \varepsilon_n) \\ E(\varepsilon_2 \varepsilon_2) & E(\varepsilon_2 \varepsilon_2) & E(\varepsilon_2 \varepsilon_n) \\ E(\varepsilon_n \varepsilon_2) & E(\varepsilon_n \varepsilon_2) & E(\varepsilon_n \varepsilon_n) \end{pmatrix} = \begin{pmatrix} \delta^2_{\varepsilon_1} & 0 & 0 \\ 0 & \delta^2_{\varepsilon_2} & 0 \\ 0 & 0 & \delta^2_{\varepsilon_n} \end{pmatrix}$$

En présence d'hétéroscédasticité, la valeur du statistique *t* et du statistique *F* seront surestimés. Nous aurons donc tendance à rejeter plus souvent qu'il ne le faudrait l'hypothèse nulle

**Test pour détecter l'hétéroscédasticité conditionnelle :**

**Test Breusch-Pagan :** Cette méthode examine si la variance estimée des résidus d'une régression dépend de la valeur des variables explicatives.

**Le test de White**

Ce test est fondé sur l'existence d'une relation significative entre le carré des résidus et les variables explicatives en niveau et au carré dans une seule équation de régression.

Une fois cette équation est estimée, l'étape suivante consiste à appliquer soit le test de significativité globale de Fisher, soit le test LM.

Plusieurs techniques de corrections peuvent être appliquées. Le principe de base de ces techniques consiste à effectuer une transformation sur les données du modèle estimé afin de rendre les écarts des résidus homoscedastique. En effet, une transformation peut s'effectuer en divisant les termes de l'équation de régression sur l'écart type du résidu.

Test d'hétéroscédasticité de White est basé sur la régression du résidu au carré sur une constante et les produits croisés entre toutes les paires de variables explicatives différentes (constante comprise). On teste l'hypothèse nulle que tous les coefficients de cette régression sont nuls, sauf le terme constant. Sous cette hypothèse, le test est distribué selon une loi Chi-2 à  $(k + 1)k/2 - 1$  degrés de liberté.

## 2- Multicolinéarité : conséquences et détection

Le terme de multicolinéarité est employé dans le cas d'un modèle incorporant des séries explicatives qui sont liées entre elles. À l'opposé, pour des séries explicatives de covariance nulle ( $Cov(x_1, x_2) = 0$ ), nous dirons qu'elles sont orthogonales. Si, pour des études théoriques, nous pouvons supposer que deux séries statistiques sont orthogonales, dans la pratique, lorsque l'économiste modélise des phénomènes économiques, les séries explicatives sont toujours plus ou moins liées entre elles.

### Les conséquences de la multicolinéarité :

- a) augmentation de la variance estimée de certains coefficients lorsque la colinéarité entre les variables explicatives augmente (le t de Student diminue) ;
- b) instabilité des estimations des coefficients des moindres carrés, des faibles fluctuations concernant les données entraînent des fortes variations des valeurs estimées des coefficients ;
- c) en cas de multicolinéarité parfaite, la matrice  $X'X$  est singulière (le déterminant est nul), l'estimation des coefficients est alors impossible et leur variance est infinie

### Tests de détection d'une multicolinéarité

Test de Klein : Le test de Klein est fondé sur la comparaison du coefficient de détermination  $R^2_y$  calculé sur le modèle à  $k$  variables :

$$y = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2 + \dots + \hat{a}_k x_k + e$$

et les coefficients de corrélation simple  $r^2_{xi, xj}$  entre les variables explicatives pour  $i \neq j$ .

Si  $R^2_y < r^2_{xi, xj}$ , il y a présomption de multicolinéarité. Il ne s'agit pas d'un test statistique au sens test d'hypothèses mais simplement d'un critère de présomption de multicolinéarité.

## 3- Problème de l'autocorrélation

Le test de Durbin et Watson (DW) permet de détecter une autocorrélation des erreurs d'ordre 1 selon la forme :

$$e_t = \rho e_{t-1} + v_t \text{ avec } v_t \rightarrow N(0, \delta_v^2)$$

Le test d'hypothèses est le suivant :

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

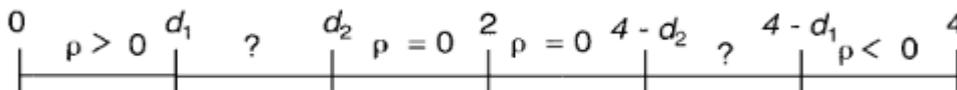
Pour tester l'hypothèse nulle  $H_0$ , la statistique de Durbin et Watson est calculée :

$$\frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

où  $e_t$  sont les résidus de l'estimation du modèle.

De par sa construction, cette statistique varie entre 0 et 4 et nous avons  $DW = 2$  lorsque  $\hat{\rho} = 0$  ( $\hat{\rho}$  est le  $\rho$  observé).

Afin de tester l'hypothèse  $H_0$ , Durbin et Watson ont tabulé les valeurs critiques de DW au seuil de 5% en fonction de la taille de l'échantillon net du nombre de variables explicatives ( $k$ ). La lecture de la table permet de déterminer deux valeurs  $d_1$  et  $d_2$  comprises entre 0 et 2 qui délimitent l'espace entre 0 et 4 selon le schéma 1 :



*Schéma 1 – Interprétation du test de Durbin et Watson*

Selon la position du DW empirique dans cet espace, nous pouvons conclure :

- $d_2 < DW < 4 - d_2$ , on accepte l'hypothèse  $H_0 \rightarrow \rho = 0$  ;
- $0 < DW < d_1$ , on rejette l'hypothèse  $H_0 \rightarrow \rho > 0$  ;
- $4 - d_1 < DW < 4$ , on rejette l'hypothèse  $H_0 \rightarrow \rho < 0$
- $d_1 < DW < d_2$  ou  $4 - d_2 < DW < 4 - d_1$  ; Nous sommes dans une zone d'indétermination, ou zone de doute, c'est-à-dire que nous ne pouvons pas conclure.

#### Conditions d'utilisation :

- le modèle **doit comporter impérativement un terme constant**;
- la variable à **expliquer ne doit pas figurer parmi les variables explicatives** (entant que variable retardée);
- pour les modèles en coupe instantanée, les observations doivent **être ordonnées** en fonction des valeurs croissantes ou décroissantes de la variable à expliquer ou d'une variable explicative soupçonnée être la cause de l'auto-corrélation ;
- le nombre d'observations doit **être supérieur ou égal à 15**.

Le test de Durbin et Watson est un test présomptif d'indépendance des erreurs du fait qu'il utilise les résidus ; de plus, il ne teste qu'une autocorrélation d'ordre 1.

Les méthodes de correction pour l'autocorrélation les plus populaires sont :

- ✓ Modèles autorégressifs
- ✓ Les moindres-carrés généralisés
- ✓ La procédure Cochrane-Orcutt